ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

AN OBJECT MEASURE EXPERIMENTAL ANALYSIS ON TWEET RUMOR IDENTIFICATION USING CLASSIFICATION ALGORITHM

Dr. A. C. Kaladevi* & S. Nithya**

* Professor, Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamilnadu

** PG Scholar, Department of Computer Science and Engineering, Sona College of Technology, Salem, Tamilnadu

Cite This Article: Dr. A. C. Kaladevi & S. Nithya, "An Object Measure Experimental Analysis on Tweet Rumor Identification Using Classification Algorithm", International Journal of Computational Research and Development, Special Issue, January, Page Number 38-42, 2017.

Abstract:

Social media analytics is a multifaceted domain. Data available on social media platforms contain diverse information abundant, and focusing on the relevant that piece of data is far from obvious and often unfeasible. Social media analytics helps to collect data as blogs and social media websites to create business decisions. The marginal effect of traditional world-of-mouth advertising is replaced by the enormous spread of information and influence through the World Wide Web. Users are no longer reluctant to share personal information about themselves, their friends, their colleagues, and their political preferences with anybody who is interested in them. Rumors can easily spread through only the social media. Nowadays, most of people share information as Twitter (Micro blog websites). People will report rumors, but rather as simply true or simply false; the majority, however simply report the fact that they have heard rumor information. Often people will hedge that repetition with warnings like "Non rumor" or "Rumor". This paper aims to find out the rumor tweets using classification algorithm in micro blog websites. Thus bag of words is used to compare tweets classification. Algorithms such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classification are object measure experimental analysis for accuracy rate, precision, recall and F- measure to find the best results. In future the work will be extended to analyze short text in rumor based tweets.

Key Words: Social Media Analytics, Micro Blog Websites, KNN (k-Nearest Neighbor) & SVM (Support Vector Machine)

1. Introduction:

Big data analytics is the use of advanced analytics techniques against very huge, diversity data sets include different types such as structured/ unstructured and different size from terabytes to petabytes. Big data is a term applied to data sets whose size to be beyond the traditional relation database and process the data with low- latency. And it has one or more following characteristics- high volume, high velocity, and high variety are shown in figure 1. Analyzing big data enables analysts, researchers and business users to make better and fast decisions using advanced analytics techniques are text analytics, machine learning, predictive analytics, data mining, statistics, and natural language processing. Data is created constantly in increasing rate. Mobile phones, social media, Image technologies create more new types of data and that must be stored somewhere for some purpose.

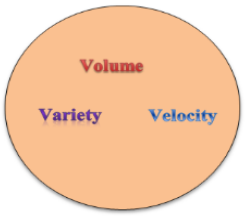

Figure 1: Big Data

Web users have been putting billions of data online on websites like Facebook (social networks sites), Twitter (micro blog websites), YouTube (multi media- sharing), Flicker (photo sharing), and LinkedIn (business-oriented social network site). Users share personal information, thoughts and opinions on micro blogs. Social media now embodies the leading and biggest source of consumer data. Social media is key model of the velocity and variety which are associated with big data. With social media, data is coming at you at an inconceivable speed and in a number of formats including videos and pictures. The worldwide popularity social media as Twitter has more than 500 million users posting and 340 million tweets per day. So that it can generate 8TB of data in each day. Micro blog websites only share information's that may be false or true data but users can't identify those information's as rumor or non-rumor.

2. Related Works:

In recent years, twitter data have become the most popular information source. Collecting Twitter data begins with identifying the topic of interest using a keyword or hash tag. While Twitter data collection can rely on standardized API services, the analysis of such collected data becomes challenging because the data are less structured. Twitter data contain a large amount of information, including tweets. The existing work explains the Hybrid Seg framework which segments tweets into meaningful phrases called segments using both global and local context. Through their framework, they demonstrate that local linguistic features are more reliable than term dependency in guiding the segmentation process. This finding opens opportunities for tools developed for formal text to be applied to tweets which are believed to be much noisier than formal text. Tweet segmentation [2] helps to preserve the semantic meaning of tweets, which subsequently benefits many downstream applications, e.g. named entity

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

recognition. The experiments, they shows that segment based named entity recognition [1] by applying segment based part-of-speech (POS) tagging methods achieves much better accuracy than the word-based alternative.

Hybrid Segmentation: HybridSeg [5] learns from both global context and local context and has ability to learn from pseudo feedback. Tweets are posted for information sharing and communication. The figure 2 explains that the global context derived from web pages therefore helps in identify the meaningful segments in tweets [6]. The local contexts are estimating the term dependency within a batch of tweets. Pseudo feedback includes the segments recognized based on local context with high confidence save as good feedback to extract more meaningful segments.

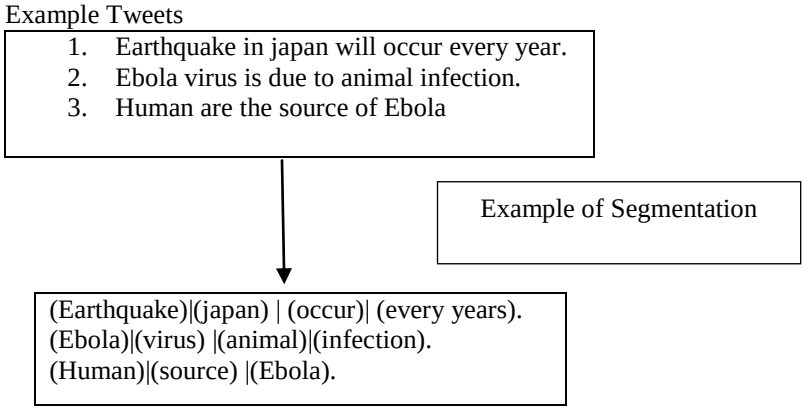

Figure 2 Hybrid Segmentation

Named Entity Recognition (NER): The figure 3 clarifies that the NER [4] is a subtask of information extraction that classify elements into pre-defined categories such as source of id, tweet id, etc.,

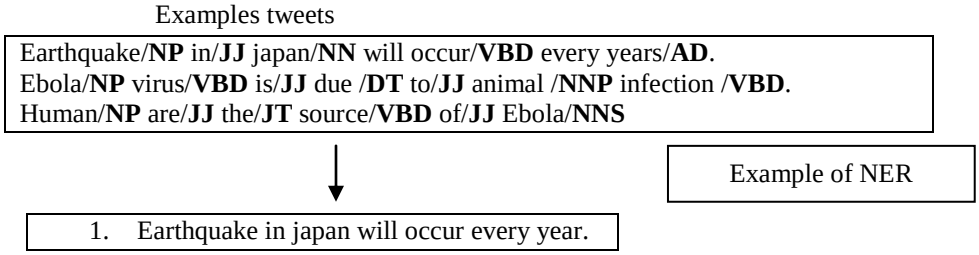

Figure 3: Named Entity Recognition

Part- of- Speech Tag (POS TAG): Figure 4 elucidates the POS tag is the process of marking a word or identification of words as Noun, verbs, adjectives, adverbs, etc.., In the POS tagger [8] to identifying the each word as a noun , verb ,adjectives , etc.., are tagging the given example tweets. Tweet segmentation is the major part for identifying a text for rumors. Many people have proposed methodologies for segmenting the tweets. Some of them are Hash tag, POS-tag, Named entity, Hybrid seg etc.

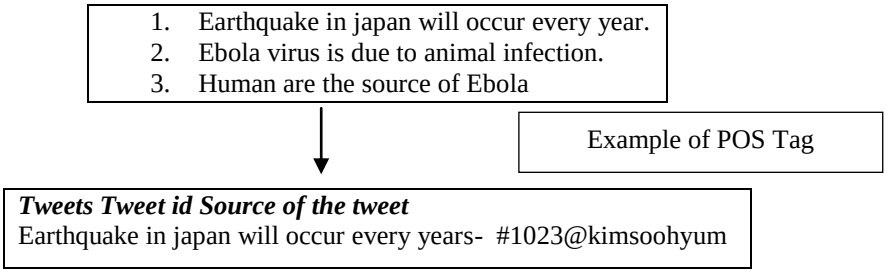

Figure 4: Part-Of- Speech Tag

3. Proposed Work:

Twitter is a micro-blogging social media platform with hundreds and millions of users. Twitter is a social network where users can publish and exchange short messages of up to 140 characters long, also known as tweets. It can define a rumor to an unverified assertion that starts from one or more sources and spreads over time from node to node in a network. Figure 5 explains the system architecture of tweet processing the big data perspective. On Twitter, a rumor is a collection of tweets, all asserting the same unverified statement (however the tweets could be, and almost assuredly, are worded differently from each other), propagating through the communications network (in this case Twitter), in a multitude of cascades. A rumor can end in three ways: it can be resolved as either true (factual), false (nonfactual) or remain unresolved. There are usually several rumors about the same topic, any number of which can be true or false. Twitter datasets are collected and stored datasets as collected in big database. The data discovery platform is used to extract the key features from uploaded datasets. The keywords are analyzed

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

based on POS tagger. Next, the analysis portfolio is used to predict the sentiments and labeled as positive or negative. It can be stored in enterprises data warehouses.

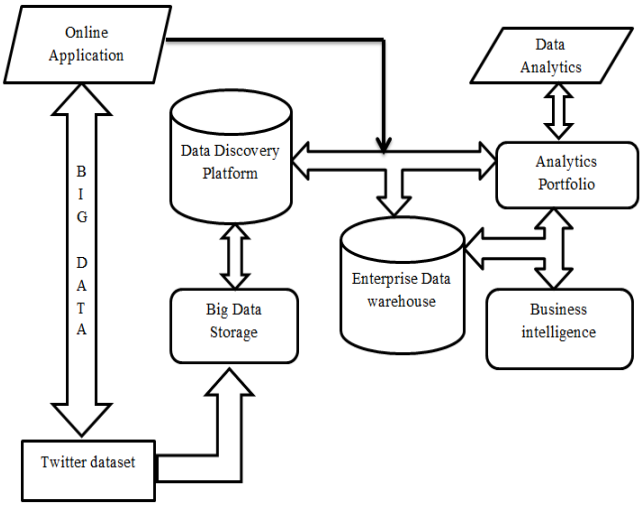

Figure 5: System Architecture

Business portfolio is used to predict the rumors based on classifiers Algorithm. The classification approach is used to label each tweets.

KNN Classifier Algorithm:

K-nearest neighbor [7] algorithm is a method for classifying words based on closest meaning to be trained using machine learning algorithm. Training process for this algorithm only consist of storing feature values and label of the training words. In the classification process, it simply is to assign the label of its k nearest neighbors. Typically the words are classified based on the labels of its K nearest neighbors. If k=1, the words are simply classified as the class of words nearest to it. They used the most common distance function for KNN which is equation (1) Euclidean distance:

Euclidean distance: $(d)^{\square} = \sum_{i=1}^{n} (xi^{\square} - yi)^2$ ---------- (1)

Symbol Explanation:
d-Distance
x,y –axis points

Algorithm: KNN Classification – Rumor Prediction

- ✓ Read the training data from a file <x, f(x)>
- ✓ Read the testing data from a file <x, f(x)>
- ✓ Set K =1
- ✓ Normalize the attribute values in the range 0 to 1
  Value=value / (1+value)
- ✓ Apply Backward Elimination
  a) For each testing example in the testing data set
  b) Find the K nearest neighbors within the training data set based on the Euclidean distance
  c) Predict the class value by finding the maximum class represented in the K nearest neighbors
  d) Calculate the accuracy as Accuracy = (# of correctly classified examples / # of testing examples) * 100

Implementation steps for KNN Algorithm:

- ✓ **Handle Data:** Open the dataset from CSV and split into test/train datasets.
- ✓ **Similarity:** Calculate the distance between two data instances.
- ✓ **Neighbors:** Locate the k most similar data instances.
- ✓ **Response:** Generate a response from a set of data instances.
- ✓ **Accuracy:** Summarize the accuracy of predictions.
- ✓ **Main:** Tie it all together.

Implementation process mainly focuses on the event of HybridSeg and KNN approach to the classification of tweets (posts on Twitter). HybridSeg learns from both global and local contexts and has the abilityto find out from pseudo feedback. In order to analyze the textual content of the tweets, give a summary of the top terms occurring in each type of topic to the classifiers. Figure 6 illustrates to collect the twitter dataset after that to apply the preprocessing methods such as filtering process to eliminate irrelevant words. The filtering process removes all the stop words contained in the tweets. The stop word removal process includes Twitter-specific words and words in stop word lists for the main languages in the dataset. Next, calculate the TF (term frequency) for each word and each type of trending topic. This process gives a list of words for each type of trending subject and ranks the words in the descending order by TF value. These steps are implemented in Global and Local Context. Before extracting pseudo feedback, POS tagger is implemented to define features in order to categorize words as adverb, adjective and so

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

on in Natural Language.Now the KNN approach is designed to classify the rumors based on three features namely content features, network features and blog features.

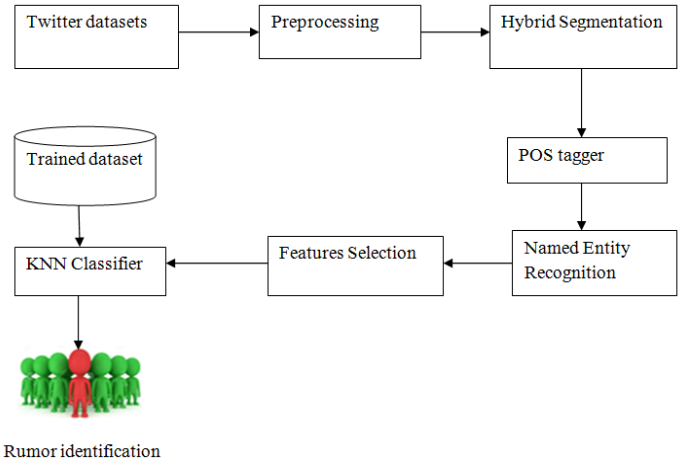

Figure 6: KNN Implementation Process

Support Vector Machine Classifier Algorithm:

Support Vector Machine (SVM) is a classifier method that performs classification tasks by constructing hyper planes in multidimensional space that separates cases of different class labels. The linear classifier that separates a set of similar meaning words into their respective groups with in a boundary line. The trained dataset to be mapped with the given input twitter dataset using set of mathematical functions known as mapping.

Linear classifier $f(x) = w^T x + b$ ---------- (2)

Such that f(x) {>=0 y_i=+1, <0 y_i= -1

Symbol Explaination:

f(x)-linear function

W^T- Normal vector

x.y –axis

b- offsethyperplane

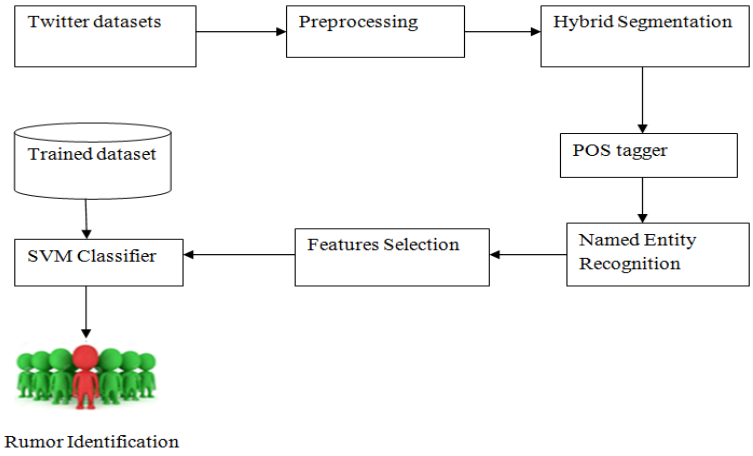

Figure 7: SVM Implementation Process

Algorithm: SVM Classification – Rumor prediction

- ✓ Read the training data from a file <x, f(x)>
- ✓ Read the Input data from a file <x, f(x)>
- ✓ Datasets are stored in separate vector
- ✓ Set the maximal margin in linear separable case
  Margin function $h(x_c)=f(x_c)-y_c$
- ✓ Choose closest meaning words from training dataset
- ✓ For a linear classifier, the training dataset is used to learn and then to classify the rumor data for given input data sets

Already describes all methodology used in above figure 7.The modification is done using SVM classifier. The SVM classifier [3] to map the training dataset and input dataset which gives similar meaning words and that word is stored in vector storage. Finally SVM classifier classifies the rumor based tweets on features selection.

ISSN: 2456 – 3137
International Journal of Computational Research and Development
Impact Factor 4.775, Special Issue, January - 2017
International Conference on Smart Approaches in Computer Science Research Arena
On 5th January 2017 Organized By
Department of Computer Science, Sri Sarada College for Women (Autonomous), Salem, Tamilnadu

4. Results and Discussion:

The table 1 clearly explains [3]to results of the KNN Classification and SVM classification based on the attributes such as Precision, Recall, F-measure. In our proposed work to classify the rumor tweets and non-rumor tweets KNN and SVM algorithms can be used. The figure 8conclude that the KNN Classification algorithm for suitable in tweets segmentation method when compare the SVM classification. The KNN Classification accuracy rate high and the time complexity and memory will be low when compare the SVM classification algorithm.

Table 4.1: Result of KNN Classification and SVM Classification

Category	K-NN			SVM		
	Precision	Recall	F-measure	Precision	Recall	F-measure
1.	0.8	0.77	0.82	0.63	0.77	0.85
2.	0.6	0.8	0.72	0.33	0.2	0.1
Average	0.7	0.785	0.77	0.48	0.485	0.475

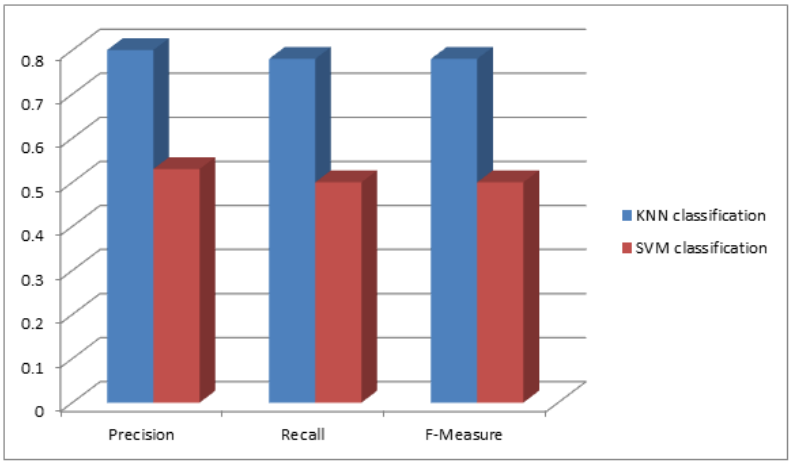

Figure 8: Performance Chart

5. Conclusion:

In this work, we implemented two different classification algorithms for tweet segmentation. We could observe that KNN classification outperformed SVM classification. Although the performance of SVM was more time complexity when compare to KNN. So the KNN classification was very effective in tweet segmentation. The KNN classification to classify the tweets based on centroid methods to check whether rumor or non- rumor. The main significance of KNN Classification to reduce time complexity and memory space when compare the SVM classification. So the KNN Classification was very effective for predicting rumor or non-rumor tweets.

6. Future Work:

The main aim to build a system that employs the findings of this work and the emerging patterns within the re-tweet network topology to find whether a new upcoming topics is a rumor or not. Future work involves using more advanced techniques from linguistics to extend the speed of correct tense identification. In specific, developments to the analysis of verb phrases and modifications of the marked parameters for sentences might be terribly useful. By creating it simple to match news coverage to twitter posts concerns a happening, the system offers both up-to-the-minute information and valuable insight into past events.

7. References:

1. A. Ritter, S. Clark, Mausam, and O. Etzioni, “Named entity recognition in tweets: An experimental study” in EMNLP, pp. 1524–1534, 2011.
2. Chenliang Li, Aixin Sun, Jianshu Weng, and Qi He, “Tweet Segmentation and its Application to Named Entity Recognition” in IEEE Transcations on Knowledge and Data Engineering,VOL.27,NO.2, 2015.
3. Jinho Kim, Byung-Soo Kim, and Silvio Savarese, “Comparing image classification methods: K-Nearest Neighbors and Support Vector Machines” in Applied mathametics in Electrical and coputer engineering. 2012.
4. Li, J. Weng, Q. He, Y. Yao, A. Datta, A. Sun, and B.-S. Lee,“Twiner: Named entity recognition in targeted twitter stream” in SIGIR, pp. 721–730, 2012.
5. Li, A. Sun, J. Weng, and Q. He,“Exploiting hybrid contexts for tweet segmentation” in SIGIR, pp. 523–532, 2013.
6. Li, A. Sun, and A. Datta,“Twevent: segment-based event detection from tweets” in CIKM, pp. 155–164, 2012.
7. S.Nithya, Dr.AC.Kaladevi, “Tweet segmentation and classification for rumor identification using KNN approach” in International conference on Big data anlaytics.
8. K. Gimpel, N. Schneider, B. O’Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith,“Part-of-speech tagging for twitter: annotation,features, and experiments” in ACL-HLT, pp 42–47, 2011.